Korean Women's Development Institute ISSUE PAPER
KWDI 이슈페이퍼
발행처 : 한국여성정책연구원
발행일 : 2023.06.15

수행과제명 인공지능 딥러닝 활용의 젠더 편향성 실태 및 개선방향
과제책임자 문미경 선임연구위원

인공지능 속 젠더 편향성 완화를 위한 정책적 개선방안

초록

◆ 챗봇 '이루다' 및 '테이'의 혐오 발언 학습 및 산출 사건 등, 알고리즘을 통해 사회에 내재되어있는 인공·젠더 차별이 발생함.
- 언론에 이슈화된 사건 외에도 인공지능 스피커의 성차별적 발언, 젠더 차별적 번역 기능, 인공지능 면접의 차별적 평가 등 일상생활에서 사용되는 인공지능 기술에서도 문제가 발생한 것으로 나타남.

◆ 따라서 인공지능 딥러닝 활용에서 발생하는 젠더 편향성 및 해당 문제 해결을 위한 대응 현황을 살펴봄. 또한 심층인터뷰 및 전문가 의견을 바탕으로 정책 방안을 제시함.
- 국내외 인공지능 젠더 편향성 사례의 유형화 및 원인 진단 결과, AI의 젠더 편향성 완화를 위해 두 가지 측면의 대응이 함께 고려되어야 함을 파악함.
- 첫째, 기술공학적 차원에서 기술 구성부터 활용까지 각 단계마다 AI의 젠더 편향성 문제를 고려하고 이를 완화하는 기술적 지침이 필요하며, 이를 본 연구에서 제시함.
- 둘째, AI의 젠더 편향성의 궁극적 원인은 사회 내 젠더 고정관념 및 성차별임. 따라서 AI 기술의 젠더 편향성 문제를 사회 전체의 문제로 인식할 필요성이 있음. 이를 위해 본 연구에서는 법·제도 개선 및 사회 전반적으로 젠더 편향성을 완화할 수 있는 사회환경 조성과 관련한 정책을 제안함.

KWDI 이슈페이퍼
2022년 한국여성정책연구원 연구보고서의 내용을 일부 발췌 및 정리하여 제작되었습니다.
문미경·김복태·강경주·김예솔·김효은·허유선(2022). 인공지능 딥러닝 활용의 젠더 편향성 실태와 개선방안. 서울: 한국여성정책연구원.

배경 및 연구 목적

- 최근 인공지능(artificial intelligence, AI) 기술이 급속도로 발전하면서 그와 관련되어 사회적 경제적 측면에서 큰 관심이 쏠리고 있음. 우리는 인공지능의 의사결정은 인간에 의한 의사결정과 달리 기계이기 때문에 중립적이고, 편향이나 편견으로부터 자유로울 것이라고 생각함. 그러나 현재의 기술 수준을 전제할 때 인공지능은 독자적 판단 주체보다 주어진 제약하에서 가능한 최적의 해답을 도출하는 통계적 분석 도구에 가까움.

- 챗봇 '이루다'의 집단 성폭력, 성희롱 학습을 통한 혐오 발언 논란, 인공지능 챗봇 '테이'가 혐오 발언을 학습하고 이를 스스로 반복 산출한 것처럼 특정한 차별 혹은 차별 기제를 학습한 인공지능이 신용거래 및 대출, 고용 후보자에 대한 평가, 대학 등 교육 기관의 입학 평가, 인공지능에 의한 개인맞춤형 기사 선별 제공, 혹은 그 외의 특정 목적을 위한 인물 선별 및 추천 검색 등에 사용될 수 있음(허유선, 2018). 알고리즘을 통해 작동하는 인공지능의 이러한 사례들은 부당한 차별적 결과와 영향이 발생할 수 있다는 것만을 보여주는 데 그치는 것이 아니라 인공지능에 의해 야기되는 차별이 인종, 젠더 등 기존 사회의 차별을 반영하며, 이를 더 강화시킬 수 있음.

- 인공지능 젠더 편향성 문제 해결을 위해서는 사람들의 젠더에 대한 편향된 인식, 이러한 인식 생산에 영향을 미치는 사회환경 등 인간과 환경의 순환적 연결고리 속에서 개선에 대한 관점이 필요함. 본 연구에서는 인공지능 딥러닝 기술의 빠른 발전 속도와 광범위한 사회적 파급력에 초점을 맞추어 인공지능 개발과정에서 젠더 편향성의 실태를 분석하고 이를 완화할 수 있는 방안을 모색함. 이를 위해 인공지능에서 야기되는 편향 사례를 성별에 초점을 맞춰 분석하였으며, 국내외 편향성을 완화하기 위한 법률과 정책 현황 등을 살펴봄. 이러한 분석에서 도출된 시사점을 바탕으로 젠더 편향성 완화를 위한 법·제도적 측면과 사회환경 구축에 필요한 개선방안과 인공지능 기술개발 현장에서 이해되고 적용될 수 있는 공학적 접근을 통한 지침 등을 도출함.

조사 및 분석결과

- 인공지능 젠더 편향성

  - 인공지능은 자율주행 자동차, 사람을 가장하는 로봇, 기계학습 등 사람들에게 각기 다른 의미를 부여하는 개념과 기술의 집합체이며 그 응용 프로그램은 어디에서나 볼 수 있음. 최근 인공지능의 정의는 인간의 지능으로 할 수 있는 문장이해, 영상인식, 음성인식, 학습 등을 컴퓨터가 실행하도록 하는 방법을 연구하는 컴퓨터 공학 및 정보기술의 한 분야로서, 컴퓨터가 인간의 지능적인 행동을 모방할 수 있도록 하는 것을 의미함. 인공지능을 통해 컴퓨터는 방대한 양의 데이터를 활용하고 학습된 지능을 사용하여 인간에 의해 소요되는 시간보다 훨씬 짧은 시간 안에 결과물을 만들어 낼 수 있음.

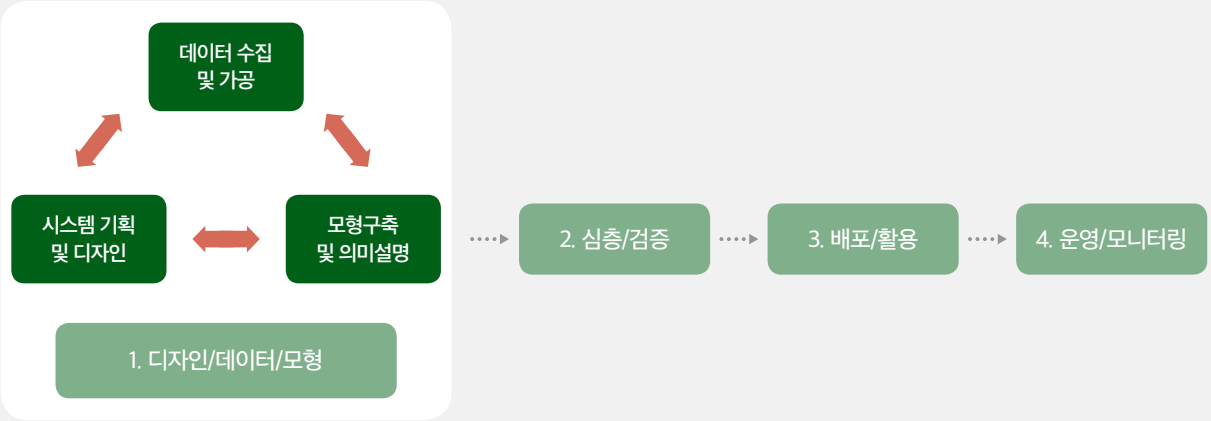
  - AI 젠더 편향성은 크게 알고리즘 설계 편향, 데이터 편향(수집, 처리, 노출 등), 인공지능의 객관성에 대한 근거 없는 믿음(인간보다 기계가 편향적이지 않을 것이라는 가정)에서 발생되는 경향을 보임. 이에 대한 대응이 쉽지 않은 까닭은 AI가 산출하는 결과물의 과정과 근거를 인간이 정확하게 이해하기도 어렵고, AI 기술의 특성(투명성 혹은 설명가능성의 문제) 때문에 일반인이 쉽게 그러한 정보에 접근하기도 쉽지 않기 때문임.

국내외 AI 젠더 편향성 사례 분석

▶ 국내외 AI 젠더 편향성 사례를 기술개발유형별로 조사하였음. 인공지능 기술 시스템은 생성 및 활용은 크게 설계와 데이터, 모델링의 기획 단계와 시스템의 검증 및 확인(verification and validation), 모델의 실제 활용을 위한 서비스화 단계인 배치(deployment), 운영 전반인 오퍼레이션(operation)과 이후 모니터링(monitoring)으로 구분됨.

▶ 국내외 사례를 분석한 결과, '인공지능 기획 및 설계 단계', '데이터 처리 단계', '알고리즘 생성 및 학습 등 모델링 단계'로 유형화됨.

<그림 1> 인공지능의 기술 구성 단계



<표 1> 인공지능 기술개발유형별 젠더 편향성 사례

인공지능 기획 및 설계 단계	
AI의 젠더화	AI 비서, 소셜 로봇 등의 여성 음성, 이미지화, 성적 괴롭힘에 대한 순응적 반응 등
성적 착취 및 수익화 목적과 AI 기술의 결합	딥페이크 기술 활용 포르노 여성 신원 추적 목적 알고리즘 개발 등
젠더 타깃 연구	얼굴 이미지만으로 성적 지향 식별가능한 안면인식 알고리즘 연구
데이터 처리 단계	
데이터 자체 편향	언어처리 영역의 기계번역, 단어 임베딩에서 성별중립 혹은 여성형 단어를 남성형으로 자동 번역 및 연관짓는 경우 자동 데이터 라벨링 혹은 자동 이미지 생성 알고리즘의 젠더에 따라 다른 결과 산출
데이터 수집 표본 집단 설정에서 젠더 편향	안면인식 프로그램의 백인 남성과 흑인 여성 간 인식률 격차 의료데이터 남성 집단 편향성
알고리즘 생성 및 학습 등 모델링 단계	
부정적인 젠더 편향 실태 및 그 개선에 대한 고려 없는 알고리즘 설계	'STEM'(Science, Technology, Engineering, Math)의 경력 개발 광고 노출의 젠더 격차 전통적 젠더 고정관념을 강화하는 컴퓨터 시각의 모델 훈련
젠더 편향 데이터 기반 알고리즘 기계학습	남성에 편중되었던 고임금 일자리 광고 노출의 젠더 격차 AI 채용 과정에서 여성 연관 키워드에 관한 무조건적 감점
알고리즘 투명성 혹은 설명가능성의 어려움	남편과 공동자산 소유자인 여성에 대하여 낮은 신용 한도 책정, 유색 인종 여성 키워드 검색 시 선정적 콘텐츠의 검색 과다 노출(검색 최상단 제시 등)

1) 기술개발과정은 AI가 기획되고 개발 및 생성되는 기술적인 측면에서의 접근, 비기술개발과정은 AI와 관련된 인권, 윤리 문제 등에 대한 접근으로 정의함

▶ 젠더 편향성 사례를 살펴본 결과, AI 기술 구성의 전체에 걸쳐 발생함. 즉, 각 단계마다 AI의 젠더 편향성 문제를 고려하고 젠더 편향성을 완화할 수 있는 유용한 도구들이 개발될 필요가 있음 또한, 실태 개선을 위한 실질적 방법의 개발 및 지속을 어렵게 만드는 근본적 원인은 역사적·구조적 제더 편향의 누적과 영향, 그리고 이를 사회의 주요 과제로 인지하지 않는 기술 업계와 사회 전반의 문제 인식으로 파악함.

국내외 AI 젠더 편향성 대응 현황

▶ 국내외 및 국제기구의 AI 관련 법률 및 정책을 기술개발과정과 비기술개발과정으로 구별하여 대응 현황을 분석함.

<그림 2> 기술개발 및 비기술개발 과정별 인공지능의 젠더 편향성 대응 현황 파악

▶ 국내 AI 관련 현행 법령 및 정책에 있어 인공지능기술 모든 단계에 있어서 차별과 편향이 발생하지 않게 할 것을 제시하고 있으나, 대부분의 법률 제정안에서 젠더에 대하여 명시적으로 고려한 조항은 미비한 상황임.

<표 2> 국내 대응 현황

법령	
기술개발과정	성차별 방지 또는 젠더 편향성 제어 조치 조항 부재
비기술개발과정	"지능정보화기본법(법률 제 17344호)"에서 포괄적 차원으로 불평등 또는 격차 유려를 포함하고 있으나, 젠더 편향성 및 성차별에 관한 명시 부재
정책	
기술개발과정	'I-Korea 4.0 실현을 위한 인공지능(AI) R&D 전략'이 있으나, 데이터 처리, 알고리즘 생성, 학습 모델링 등 세부 개발과정에 대한 정책적 접근은 부족, 정책 전반에서 젠더 요소에 대한 고려 부재
비기술개발과정	젠더 관련 내용 부재
의원입법 법률	
기술개발과정	'알고리즘 및 인공지능에 관한 법률' 성별 등에 의한 차별대우 배제를 명시화
비기술개발과정	'인공지능 연구개발 및 산업 진흥, 윤리적 책임 등에 관한 법률' 제정안에서 인권과 존엄성 보호 규정이 명시되어 있으나, 성평등적 관점 부재

<표 2> 국내 대응 현황

정책 가이드라인	
기술개발과정	과학기술정보통신부 '지능정보사회 윤리 가이드라인 국가인권위원회 '인공지능 개발과 활용에 대한 인권 가이드라인 한국여성민우회 '페미니스트가 함께 만드는 AI 가이드라인
비기술개발과정	국가인권위원회 '인공지능 개발과 활용에 대한 인권 가이드라인 한국여성민우회 '페미니스트가 함께 만드는 AI 가이드라인

▶ OECD, EU, UNESCO 등의 국제기구들은 인공지능 윤리에 대한 권고안을 마련하여, 각 국가들이 AI 기술을 개발하고 활용하는 데 있어 신뢰성을 확보하고 AI 윤리를 고려해야 함을 강조함. 해외 주요 국가들의 법제화 선례와 같이 AI가 인종·성 차별에 미칠 수 있는 파급효과에 대하여 투명한 방식으로 평가 및 공개하는 방안에 젠더 편향성을 포함시킬 필요가 있음.

▶ 우리나라 인공지능 기술 근거를 이루는 지능정보화 기본법 속에 젠더 편향성을 완화할 수 있는 내용을 포함할 필요가 있음. 또한, 유네스코 AI 윤리 권고에서 시사점을 얻어 윤리영향 평가와 모니터링 역할을 수행할 전담기구를 설치하고 운영할 때 젠더 편향성을 완화할 수 있는 내용들을 포함하여 구성할 필요가 있음.

심층 인터뷰

▶ 인공지능 개발 경험자 10명(남성 6명, 여성 4명) 대상으로 심층 인터뷰 진행. 인터뷰는 설문을 통한 조사(1단계), 사회문화적 환경에 대한 심층 인터뷰(2단계), 편향완화 기술에 대한 역량 및 인식 현황 조사(3단계)로 진행함.

▶ 인공지능 젠더 편향성은 데이터 수집 및 선정 단계, 데이터 처리 단계에서 충분히 주의를 기울이지 않은 점이 주요 원인으로 파악됨. 이는 인공지능 구성 방향과 관련한 최종 의사결정자들(관리자)의 근본적인 책임이 있음. 다만 개발과정의 시간 및 비용의 한계상 편향성 완화 기술과 요소를 적용하기 어렵기 때문에, 기술적 차원의 젠더 편향성을 완화할 수 있는 지침 등을 개발할 필요가 있음.

▶ 현재 AI 연구개발 현장의 30대 개발자 중 여성 비율이 현저히 낮다는 점과 성별 자체보다는 성차별 경험이 개발과정에 영향을 미칠 수 있다고 인식하고 있음. 특히, 첨단을 다루는 과학기술분야에서 성차별적 근로환경은 여성들의 경력을 단절시키는 주요 요인으로 작용할 수 있음.

▶ 심층 인터뷰를 통해 1) 인공지능 윤리 젠더 편향성 체크리스트, 2) 공학자를 위한 인공지능 기술개발 단계 전데 편향성 완화 지침, 3) 개발자를 위한 인공지능 기술별 젠더 편향성 완화 지침과 성별 균형 인력양성 등 사회환경 구축에 필요한 제언을 도출함.

정책제언

◉ 인공지능 젠더 편향성의 국내외 주요 사례의 유형화 및 원인 진단에 따르면 AI의 젠더 편향성의 완화를 위해 두 가지 측면의 대응이 함께 고려되어야 함을 파악함.

▶ 1. 기술공학적 차원에서 기술 구성 및 활용 전체 단계의 젠더 편향성 문제를 완화하기 위한 대응이 필요함. 앞서 AI의 젠더 편향성 문제는 AI 기술 구성의 단계별 발생할 수 있음을 확인함. 따라서 AI 기획 및 설계, 데이터 처리(수집·가공·관리 등), 알고리즘 생성 및 학습 등의 모델링이라는 AI 기술 구성의 전체에 걸쳐, 각 단계마다 AI의 젠더 편향성 문제를 고려하고 이를 완화하는 기술적 지침 등이 필요함.

▶ 2. AI의 젠더 편향성의 궁극적 원인은 사회 내 젠더 고정관념 및 성차별임. 따라서 AI 기술의 젠더 편향성 문제를 사회 전체의 문제로 인지하고, 국제기구에서 논의되는 인공지능 관련 법제와 정책의 시사점을 통해 법·제도 개선 및 사회 전반적으로 젠더 편향성을 완화할 수 있는 사회환경 조성과 관련한 정책들을 제안함.

<표 3> 개발과정유형별 정책적 제언

1. 기술개발과정에서 개선방안	인공지능 기술 및 시스템 구축단계
	(1) 인공지능(AI) 윤리 젠더 편향성 체크리스트 개발 및 보급 (2) 공학자를 위한 인공지능 기술개발 단계 젠더 편향성 완화 지침 개발 및 보급 (3) 개발 기술자를 위한 인공지능 기술별 젠더 편향성 완화 지침 개발 및 보급
2. 비기술개발과정에서 개선방안	법·제도적 개선
	(1) 지능정보화기본법(제3조) 기본원칙에 젠더 고려 명시 (2) 지능정보화기본법(제56조) 젠더 영향평가 명시 (3) 인공지능 윤리영향평가(AI Ethical Impact Assessment)에 젠더 편향성 금지 항목 포함 (4) 성인지 관점에서 AI윤리심의기구 구성 및 운영 (5) 정부차원의 AI젠더 편향성 모니터링단 운영 (6) 데이터 및 알고리즘 관련 젠더 편향성 검증기준 법제화
	사회환경 조성
	(1) 인공지능 윤리 및 젠더 편향성 교육 실시 (2) AI 기술 및 기초과학 분야 여성 인재 양성 및 경력 개발 지원

참고자료

네이버 지식백과 인공지능

(https://terms.naver.com/entry.naver?docId=1136027&cid=40942&categoryId=32845, 접근일 2022.4.12.)

뇌와 인공지능의 알고리즘.

(https://misterio.tistory.com/entry/%EB%87%8C%EC%99%80-%EC%9D%B8%EA%B3%B5%EC%A7%80%EB%8A%A5AI%EC%9D%98-%EC%95%8C%EA%B3%A0%EB%A6%AC%EC%A6%98Algorithm, 접근일 2022.03.17.)

양종모(2017). 인공지능 알고리즘의 편향성, 불투명성이 법적 의사결정에 미치는 영향 및 규율 방안. 법조 2017-6.

허유선(2018). 인공지능에 의한 차별과 그 책임 논의를 위한 예비적 고찰 -알고리즘의 편향성 학습과 인간 행위자를 중심으로-. 한국여성철학 29.

OECD (2019:15). Artificial Intelligence in Society. Paris: OECD Publishing.

주관부처 : 과학기술정보통신부 인공지능기반정책과, 과학기술정보통신부 정보통신정책총괄과, 여성가족부 여성인력개발과, 여성가족부 권익침해방지과

관계부처 : 과학기술정보통신부 인공지능기반정책과, 과학기술정보통신부 정보통신정책총괄과, 여성가족부 여성인력개발과, 여성가족부 권익침해방지과